ENHANCING PERFORMANCE OF TRAFFIC CLASSIFICATION WITH FEATURE SELECTION METHODS

Htay Htay Yi

Associate Professor, Information and Communication Technology Research Center, Information and Communication Technology Training Institute, Hlaing Campus,
Yangon, 11051, Myanmar
htayhtayyee@ictresearch.edu.mm
http://ictresearch.edu.mm

Khaing Khaing Wai

Professor, University of Computer Studies, Yangon, Department of Information Technology Support and Maintenance, Shwe Pyi Thar Township,
Yangon, 11411, Myanmar
khingkhingwai@ucsy.edu.mm
http://ucsy.edu.mm

Abstract

Network security is to protect conscious data, and some people do not know how to protect sensitive data. The system is correct the firewall rules that consistent within the organization. This work created a comprehensive testbed architecture that combined a firewall with an IDS to generate a dataset using usual traffic, DoS attacks, and Port Scan attacks. Feature selection methods like Correlation-based Feature Subset (CS) and Correlation Attribute (CA) can help reduce the complexity and conserve system resources in the context of network security and intrusion detection. The proposed dataset compared with CICIDS2017 that the performance improves without considering the flag features. The performance calculated with the CS method and compare with and without considering flag features in CICIDS2017. When using the CA Method, the minimum boundary value is determined by taking the average value of the two datasets based on the trains of the features. It finds the good features that extract based on the destination host of the desired traffic. The system contrasts the reduction of unnecessary attributes in both the proposed and CICIDS2017 datasets to enhance the adequacy of performance, especially as a False-Positive Rate (FPR) and accuracy between them.

***Keywords*: Network Security; False-Positive Rate; Performance; Feature Selection Methods.**

1. Introduction

With more people using the internet, there are more potential targets for cybercriminals. It means more opportunities for attacks on individuals, businesses, and organizations. The proposed system implements a network design that includes the firewall, and Intrusion Detection System (IDS) that analyzes DoS and Portscan traffics and applies machine learning to detect intrusion. The feature reduction or selection methods are applied for the DoS/DDoS attack traffics and illustration techniques to asset the appropriate methods for detection by researchers [Bouzoubaa (2021)], [Kshirsagar (2021)]. As the rules in the firewall work, as usual, the rules are out of order, and the administrator's typing error may be the weakness of this system [Yi (2019)]. Depending on the large organization, the administrator's mistakes in firewall configuration will cause the network to become a weakness, and the system will use a tool that is not difficult to use. In addition, they solve to reduce the number of firewall anomalies by using the Novel-Rule relations model [Valenza (2020)] and a rule merging algorithm depending on the service [Zhang (2015)]. It reduces the anomalies between the firewall rules that help the administrators to reduce the manual update rules. The proposed system is small and since the number of rules increases the firewall's functions, it can affect the network performance more or less, so it does not set many rules using algorithms, but only the rules that match the organization are manually tested and set for the correctness of each rule.

There are two main patterns in IDS based on signature and based on anomaly. If a signature-based attack comes in, the attack can detect by using predefined rules. Anomaly-based is a statistical pattern, if there is a change that doesn't match that pattern, it can be known as an attack. This work relates with Snort, an open source Intrusion Detection System that examines and detects as protocol and content. The data pre-processing included: data consolidation, data evolution, data depletion, and data cleaning before determining whether to improve the overall

pattern and take the time. The depletion of attributes applied feature subset selected methods to reduce the number of attributes [Pervez (2014)]. The authors of the research papers have done with the CICIDS2017 dataset and used the selected feature method to divide the features into groups and calculate the performance [Kurniabudi (2019)]. When setting feature groups, it is possible to group the weak features, and on the other hand, group the strong features, and the resulting performances may be different. In comparison with 26 features of CICIDS2017 [Yi (2021)], the feature selection method is not considered, and its requirement is examined in this paper.

In the proposed system, when considering the instances in the dataset, it is added based on the inbound and outbound traffic of the destination host rather than the source host and destination host. The feature selection method is crucial to enhance the performance of the system by reducing the unnecessary features in pre-processing of data. In this work, the presentation of the system and the quality of the features obtained using the selected feature methods calculate with classifiers. By using sixteen features in the proposed system and calculating the performance obtained using the Correlation-based feature subset (CS) selection method and Correlation attribute selection (CA), the user can know the goodness of the features. The main research area is:

- Assigned the firewall rules on the five interfaces with network services.
- Implemented of IDS with predefined rules and verifying with machine learning classifiers.
- Proposed the dataset with the sixteen features that are relevant to DoS and PortScan attacks to accomplished the proposed dataset to enhance the performance.
- Proven that removing the flag features do not affect the performance of the system with feature selection methods.
- Compared the proposed and CICIDS2017 datasets to prove the effective features and superior achievement of performance.

This work is collected as follows. Section 2 covers the principal of the material, method of the system. Section 3 works of the previous authors. Section 4 provides the system setup and design. Section 5 proves the result and discussion of the system with proposed dataset and compares with existing dataset. Section 6 is the conclusion of this work.

2. Methodology of the system

This section will discuss the importance of firewalls for security and why the IPCoP firewall chose in the system. In addition, Snort, which is open source from IDS, will continue to be presented. The machine learning classifiers, and existing datasets related to the proposed system will present from the content written by previous.

2.1. *Firewall*

Firewalls are a prevailing technology that are examined for security matters. In order to protect against the attack of unwanted intruders between subnetworks, policies are set in the firewall. A firewall can set policies and protect against unwanted intruder attacks between subnetworks, and this policy assigns in the firewall filtering field. These fields are network fields, protocol type, Source, and Destination (IP address and port), which are present with the action field. It is selected based on three factors, the first factor is features, the second factor is the function of the firewall required for the organization, and the third factor is selected based on the budgets the organization can use. This system uses an open-source IPCoP firewall, has good features, and is without paying.

2.2. *IPCop Firewall*

There are many software-based firewalls, among them IPCoP firewall using IPCop 2.1.8, which is adequate to grant the installation and add the packages if needed to build up in firewalls. This is a Linux Firewall Distribution that keeps up the fixed, and secure. When setting policy in the IPCoP firewall, it sets based on the service needed by the organization. It has an add-on feature, so if the users need to add more packages, they can easily add them. The IPCoP has collected four network interfaces that are Green, Red, Blue, and Orange in Fig 1. IPCop implements in five web interfaces. The first is outgoing traffic, the second is IPCop access, the third is internal traffic, the fourth is external IPCop access, and the last is port forwarding. When setting allow or deny rules on four interfaces, it can be set based on default rules.

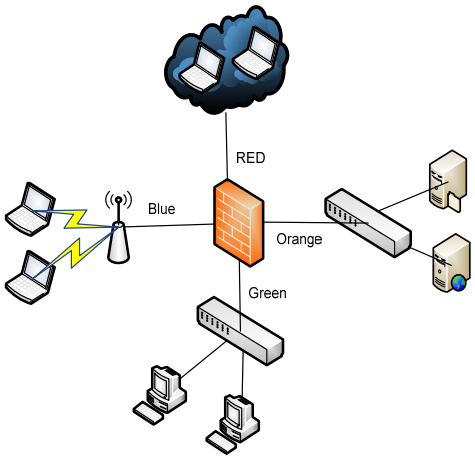

Fig. 1. IPCoP Firewall Network Interfaces.

2.3. *Snort IDS*

An intrusion detection system (IDS) observes that the service is agreeable to predefined rules. If communication is found matching a rule, the system judges a critical event related to invasion and reveal an alert to the network administrator [Yi (2019)]. IDS has three modes, the first is sniffer mode, the second is packet logger mode, and the last one is intrusion detection mode. The proposed system now uses intrusion detection mode to support the detection. Snort rules typically consist of two main parts: the rule header and the rule options. The rule header contains information such as the action to take when a packet matches the rule, the protocol to inspect, and the source and destination IP addresses and ports. The rule options specify additional criteria for matching packets, such as content to search for within packets or specific flags in TCP headers. The rule format that shows below.

Rule Header - alert tcp any HOME_NET -> 192.168.56.50/32 443
Rule Options - (msg: "HTTPS access from anywhere", sid=1000009;)

The general form of a Snort rule:
action proto src_ip src_port direction dst_ip dst_post (option)

Actions: The “log” action instructs Snort to log information about the packet that triggered the rule to a log file. It is typically used for passive logging and doesn't generate alerts or notifications. The “alert” action creates an alert when a rule matched, indicating a potential security event. When Snort encounters a rule with the "alert" action, it generates an alert message and may perform additional actions based on the configured alerting method.
Protocols: The field is operated to specify the network protocols to which the rule applies. The values are indicated IP, ICMP, TCP, and UDP.
IP addresses: It defined IP addresses, source and destination, and along ports are integral for defining the rule.
Ports: The port field allows the user to define which ports the rule should match against. The port field accepts single ports as port ranges, and the syntax for specifying port ranges involves using a colon to separate the upper and lower bounds of a range.
Options: It is associated with Snort plugins, which are modules responsible for performing various types of analysis on network traffic. When a rule is matched against a packet, Snort processes the associated options using the corresponding plugins to perform additional scanning or analysis on the packet.

3. Literature Review

In this section, Firewall, IDS, machine learning classifiers, and existing datasets related to the proposed system will present from the content written by previous authors.

3.1. *Awareness of Firewall and IDS*

The researchers accommodate with effective NIDSs and Firewalls. In [Liao (2013)], it described the detection methods, perspective, and awareness of IDSs. The researchers acquainted with studies and open-source tools that

they learned of IDS. In [Alhomoud (2011)], Network Intrusion Detection Systems (NIDS) are applied to join the technique. They take aside one to decide the benefit and the effect of Snort, and the other is Suricata. The network attacks is presented in [Bijone (2016)]. As benefit two popular intrusion detection systems (IDS), focusing on their impact in network security, based on the classification.

3.2. *Concern with Machine Learning Classifier*

Weka project team. C4.5 developed the J48 and it is an addition of ID3 algorithm. The performance of accuracy applied the J48 in anomaly detection [Aljawarneh (2017)]. Most machine learning models have over-fitting problems when constructing them. So, many researchers used k-fold cross-validation to prevent this problem.

3.3. *Feature Selection Methods with Existing Dataset*

Most machine learning models have over-fitting problems when constructing them. So, many researchers used k-fold cross-validation to prevent this problem. In [Mukkamala (2016)], neural network and Support Vector Machine (SVM) applied the machine learning approach with a 1998 DARPA dataset. The two classifiers are compared with the intrusion detection system. The authors reviewed the previous papers based on feature correlation, time consumption, and performance that were evaluated, provided, and accessed [Mauro (2021)]. It can see that a lot of resources use for good performance, and if one is good, one can be a weakness. In [Pervez (2014)], Machine-Learning and Data-Mining techniques are applied in CIC-IDS2017 and CSE-CIC-IDS2018 to review. In [Kshirsagar (2021)], a novel technique improves the DoS attack detection rate through three feature-selection methods. The existing datasets that CICIDS2017 and KDD-CUP-99 used three filter-based feature reduction algorithms. The Wrapper-based feature selection is applied, and the performance of the detection rate is with the baseline method [Albarka (2020)]. Using a supervised-machine-learning algorithm with four datasets, it surveys and prove that feature selection improves in calculating performance [Abdallah (2022)]. The successful results of using three machine learning methods, SVM, KNN, and Decision Tree, are described on five datasets that grew the IDS system extensively [Kilincer (2010)]. The author demonstrates the good results using two datasets, KDD99 and DARPA 1999, with the correlation-reduction feature selection method [Kamarudin (2019)]. The author proposed hybrid feature selection (combine filter and wrapper methods) using a random forest (RF) classifier and showed the advantage of reducing features. There can be studies on the usefulness and accuracy of the proposed method on Anomaly-based-IDS [Maseer (2021)]. In addition, the performance of different Machine Learning Anomaly-based-IDS proved in web attacks.

In [Panigrahi (2021)], how many classifiers applied from the literature status, performance, and paper gaps have been reviewed in the IDS field? Also, from this review, there can see how well the J48 classifier works with IDS. Weka project team. C4.5 developed the J48 and it is an addition of ID3 algorithm. The performance of accuracy applied the J48 in anomaly detection [Aljawarneh (2017)]. The paper [Mohammadi (2021)] observed that the detection performance is good by using an SVM classifier on IDS. It supports a repository of ML-IDS that is easy to implement and understand for researchers using conventional and ultra-modern networks [Yang (2022)]. By building a threshold model to help detection in advance and testing it with three machine learning classifiers, this can learn the most suitable classifier with the highest detection rate [Tobi (2019)]. In the previous paper, there can learn from reviewing the usefulness of ML Classifiers using a Web Application Firewall (WAF) and using signature-based attack patterns to prevent web attacks [Applebaum (2021)]. The proposed system described the false positive rate and accuracy of the two feature selection methods are a performance improvement.

4. Proposed System Setup

The proposed system design operates software-based firewall as IPCop. In Fig 2, the firewall is composed of Wide Area Network (WAN), Local Area Network (LAN) and, De-Militarized Zone (DMZ) for public and local user's access. For the LAN network, De-Militarized Zone (DMZ) is combined as an additional security layer. The local and public users can access the web server and file server in DMZ.

In Firewall, the main three zones are defined the rules with compatible for our organizational users. The forwarding rules are needed for public user access in the web server and file server in the DMZ. When setting firewall rules, not only good security but also the system performance is taken into consideration. The predefined rules of Intrusion Detection are concerned with firewalls. The firewall policies for users in the organization; and external users; It is set so that the network performance of the system is not reduced. The firewall policies are as follows:

(1) External users allow to ping and access Web server and File server in the DMZ but not through remote access (for example ssh service); Do not grant all access to the firewall.
(2) Internal users allow to ping the Firewall and DMZ network, but remote access and https services are only given to an administrator. In the DMZ, http and ftp services are allowed to all local users including the

administrator. If local users do not follow the rules established by the organization, internet access is blocked and a rule is set.

(3) When setting rules for the firewall, only the administrator is allowed to manage the firewall using https with Web access and the ssh with remote access.

When setting a rule in the firewall, the rules that match the organization are set based on the firewall default rules. The web interfaces of the firewall set individual rules for three rules for External IPCop Access, five rules for Port Forwarding, seven rules for Internal Traffic, eight rules for IPCoP Access, and three rules for Outgoing traffic, with a total of 26 rules based on firewall policies.

The Intrusion Detection System is utilized two NIC cards, is allowed external and internal users to access web and file servers in the DMZ, and has an administrator IP address for ssh access. By integrating the firewall and IDS, it can be found out about attacks that are not known by the firewall in the alerts of IDS. In addition, it makes the system secured and does not affect network performance.

The system design is accomplished two ubuntu 20.04 machines that used attackers in public network. The web and ftp server utilized with OpenSuSE 15.1 in the DMZ network implementation. The administrator and local user PCs are setup with OpenSuSE 13.2 operating system. The implementation of the web server that the services as DNS, HTTP, and SSH are installed.

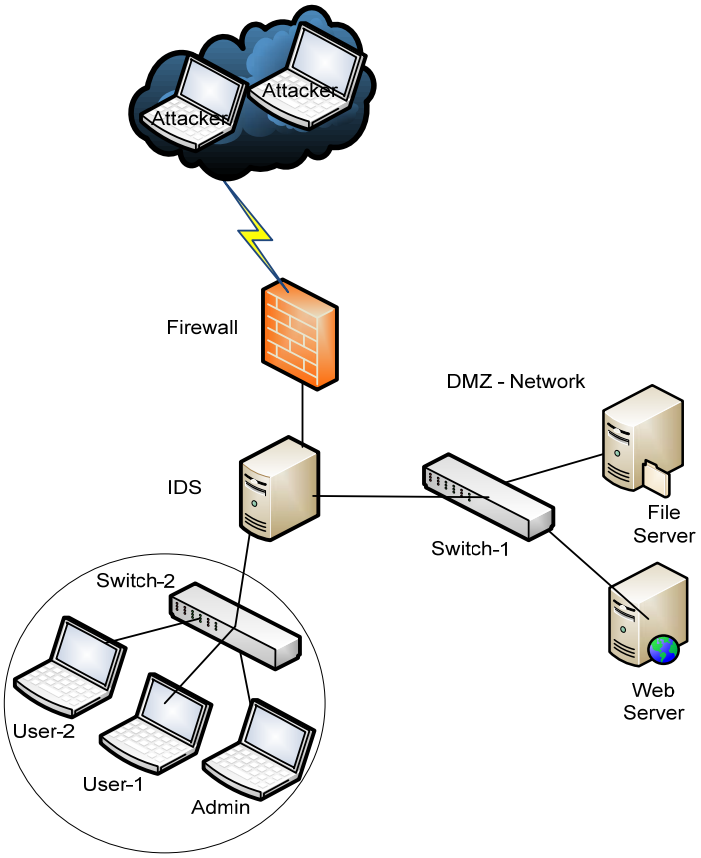

Fig. 2. Proposed Network Design

4.1. *Dataset with Network Traffic*

To implement dataset, we create DoS and PortScan attacks by using hping3 tool. The Nmap tool is used for network traffics between the external network and the web server in DMZ. For the normal traffic, it used Google, Amazon, Facebook and, YouTube to the public access. The normal traffic and attack traffic are taken by using the TcpDump tool, and the packets are taken from it, and it becomes a pcap file. Using the data in the pcap file, we use the Wireshark tool to select features and set their values.

Synchronous, Synchronous-Acknowledgement, retransmission, and reset are subdivided into smaller pcap file according to the packet range. The comma-specified file (csv) implements that aggregates values of good features. The value of features calculated that it filters out on the destination host IP address in the packet range. The normal traffic and attack traffic capture on different time that create on the report of instances weight and package range from the different traffics of csv file become a proposed dataset.

4.2. *Selected Features for Proposed Dataset*

In Table 1, the proposed dataset is composed of sixteen keys features. The package that synchronous, acknowledgment, retransmission, and reset packets are categorized into package ranges. In building the proposed dataset, some features were acquired from CICIDS2017 as an example of destination port and minimum and maximum packet length [Kurniabudi (2019)], [Thakkar (2020]. The flag features do not significantly change performance, so all are not considered in the proposed dataset. Features considered depend on the input and output of the destination host. They are destination inbound/outbound packet, total Input/Out packet, etc.

No	Features	Description
1	Dst_port	Destination Port
2	Dst_IP	Target IP Address
3	Total_Inpkt	Total Inbound packages to destination host
4	Total_Outpkt	Total Outbound packages from destination host
5	Inpkt_bytes	Inbound packages bytes to destination
6	Outpkt_bytes	Outbound package bytes from destination
7	Total_InOut_pkt	Total packages to/from destination host
8	Inpkt_bits/s	Inbound packet bits/s to destination
9	Outpkt_bits/s	Outbound packet bits/s from destination
10	Protocol	Protocol as TCP or UDP
11	Service	Service type as http, ftp
12	Min_pktlen	Minimum packet length in the packet range
13	Max_pktlen	Maximum packet length in the packet range
14	Avg_pktlen	Average packet length that fall in the packet range
15	InOut_count	Number of packet count with source and destination IP in this range
16	Class	Describe normal or attack

Table 1. Features of dataset from traffics.

4.3. *System Block Diagram*

In block diagram, we create the rulesets from the system services as shown in Fig 3. The predefined rules applied in the Firewall and IDS. When a packet enters the firewall, it is checked against the predefined rules and allowed if it matches the rules and not allowed if it doesn’t compare the predefined rules. The firewall will check the packet again with the rule specified by IDS. The firewall doesn't know the attacks by intruders that can be known through the logs due to the rules defined by IDS. The dataset is implemented based on the network traffic that defined by rules in the testbed.

Machine learning is used to verify of the values and instances of the dataset are valid. Using machine learning classifiers, the performance of the proposed dataset has been proven to be good in terms of false positive rate and accuracy. In addition, the feature selection method of the dataset, correlation-based feature selection and gain ratio, has been used to select the features and then machine learning classifiers have been applied to further prove the performance. The proposed dataset will be compared with the existing dataset CICIDS-2017 to prove the good performance as low false positive rate and high accuracy with full features and selected feature CFS.

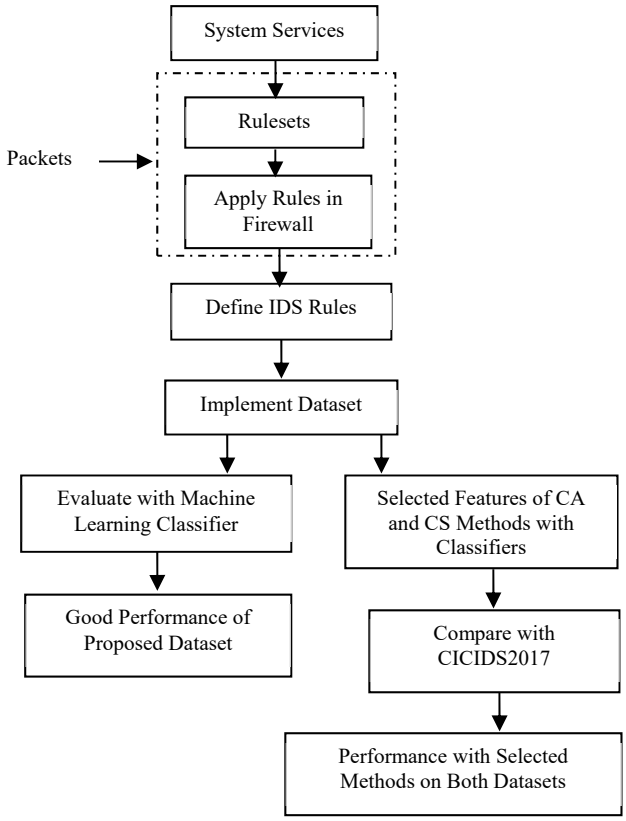

Fig. 3. Proposed Network Design.

5. Proposed of the System Results

The proposed system has done sixteen constructive features with WEKA (Waikato Environment for Knowledge Analysis) tool to show the superior features and their values with machine learning methods. The 10-fold cross-validation used as a validation method. This implementation concerns six machine learning classifiers. There are Logistic Regression (LG), SVM (Support Vector Machine), Naïve Bayes (NB), Bayes Net (BN), Random Tree (RT), and J48. When the false-positive rate is high, the actual attack cannot be detected, so we will represent the false-positive rate and accuracy of the system performance.

$$False\ Positive\ Rate\ (FPR) = \frac{Incorrectly\ Normal\ Classified\ Instances}{Total\ Normal\ Instances} \quad (1)$$

$$\text{Accuracy} = \frac{\text{Correctly Classified Instances}}{\text{Total Number of Instances}} \quad (2)$$

In Table 2, the machine learning classifiers of J48 and Random Tree (RT) have the maximum false positive rate that are 2.3% in PortScan attack. The maximum false positive rate is only 1.3% at Random Tree (RT), and the other classifiers is reduced the false positive rate. The system accuracy is calculated based on the following equation. It can see that in terms of accuracy, the lowest for DoS is approximately 98.8% in Random Tree (RN), while PortScan has also 98.8% in Naïve Bayes (NB). The accuracy of the remaining classifiers is above 99% in both attacks. It demonstrated that the false-positive rate (FPR) and accuracy (Acc) are significantly better in a proposed dataset.

Detection Classifiers	PortScan Attack		DoS Attack	
	FPR	ACC	FPR	ACC
LG	0%	99.1%	0.5%	99.5%
SVM	0%	99.54%	0.7%	99.6%
NB	0%	98.86%	0.4%	99.9%

BN	0%	99.54%	0.1%	99.9%
J48	2.3%	99.77%	0.1%	99.9%
RT	2.3%	99.1%	1.3%	99.8%

Table 2. Performance with Classifiers in DoS and PortScan Attack.

5.1. *Feature Selection Method with Proposed Dataset*

The proposed dataset operates Correlation-based feature subset (CS) to demonstrate the implicit features. The CS selects the proposed system of the subset of six selected features in DoS and five features in PortScan attacks. The increasing number of false-positive rates impacts on the detection rate of IDS. So, this section will introduce the false-positive rate.

Classifiers	Selected Features (PScan)	Selected Features (DoS)	FPR (DoS)	FPR (PScan)	Acc (DoS)	Acc (PScan)
LG	Outpkt_bits/s, Services, Min_pktlen, Max_pktlen, InOut_count	Dst_port, Inpkt_bits/s, Min_pktlen, Max_pktlen, Avg_pktlen, InOut_count	0.1%	2.3%	99.9%	99.54%
SVM			2.2%	47.8%	98.21%	94.51%
NB			0.4%	0%	99.5%	99.54%
BN			0%	0%	99.99%	99.99%
J48			0.1%	0%	99.9%	99.99%
RT			0%	0%	99.99%	99.99%

Table 3. Performance with CS Method with Proposed Dataset

The Table 3 proves the low false positive rate with machine learning classifiers except in 47.8% for SVM and 2.3% for Logistic Regression (LG) classifiers in PortScan attack. In the case of DoS attacks, it can see a significant decrease in the false-positive rate. It proves that the maximum false positive rate is only 2.2%. Calculating the accuracy of the proposed system by using CS shows that the DoS attack has at least 98%. It proves that the SVM classifier and the accuracy of the PortScan attack is 94.5%, while the rest of the classifiers are higher.

5.2. *Existing Dataset CICIDS2017*

In CICIDS2017 dataset, it contains a massive of traffic and 78 features to be detected for anomalies [Pervez (2014)]. It included two types of traffic attacks and normal traffics improved the detection rate of IDS [Thakkar (2020)]. CICIDS-2017 is involved seven attack types [Kamarudin (2019)]. It is presented using a correlation-based feature selection method (CS) from all 78 features and the obtained features and proves the system performance in Table 4 (a).

Detection Classifiers	Features (PScan)	Features (DoS)	FPR (DoS)	FPR (PScan)	Acc (DoS)	Acc (PScan)
LG	Bwd Packet Length Mean, PSH Flag Count, Init_Win_bytes_backward, act_data_pkt_fwd, min_seg_size_forward	Destination Port, Total Length of Bwd Packets, Init_Win_bytes_forward, Idle Max	27.3%	0.8%	91.21%	99.25%
NB			2.6%	3.4%	65.6%	99.1%
BN			0.9%	0.7%	64.2%	99.32%
J48			0.8%	0.2%	99.77%	99.83%
RT			0.8%	0.2%	99.77%	99.84%

Table 4. (a) Performance with CS Method with CICIDS2017 Dataset (with Flag)

Detection Classifiers	Features (PScan)	Features (DoS)	FPR (DoS)	FPR (PScan)	Acc (DoS)	Acc (PScan)
LG	Bwd Packet Length Mean, Init_Win_bytes_backward, act_data_pkt_fwd,	Destination Port, Total Length of Bwd Packets,	27.3%	10.9%	91.21%	99.1%
NB			2.6%	42.9%	65.6%	98.86%
BN			0.9%	0.4%	64.2%	99.54%

J48	min_seg_size_forward	Init_Win_bytes_forward, Idle Max	0.8%	0.2%	99.77%	99.77%
RT			0.8%	0.2%	99.77%	99.1%

Table 4. (b) Performance with CS Method with CICIDS2017 Dataset (without Flag)

The performance of five machine learning classifiers is presented based on the features derived from CS, which are four features in DoS and five features in PortScan. In the PortScan attack, the maximum false-positive rate is 3.4% at Naïve Bayes (NB), and Other classifiers found to be significantly less in Table 4 (a). In the DoS attack, the Logistic classifier has a significantly higher false positive rate, but the rest of the classifiers have a lower false positive rate.

By removing 13 flag features from the 77 features of CICIDS2017 and calculating the performance based on 64 features, it found no change in DoS attacks and only small false-positive rate (FPR) changes in the PortScan, as seen in Table 4 (b). PortScan does not include the PSH Flag Count feature and the resulting outcome have no significant changes in the classifiers, except for the Logistic Regression (LG) and Naive Bayes (NB) classifiers.

Table 4 (a) and (b) show the performance as accuracy obtained when considering and not considering the thirteen flag features of CICIDS2017. The accuracy obtained when flag features are not considered is 98.86% in Naive Bayes (NB), but the performance of the other classifiers is not affected. Especially in a DoS attack, it can be clearly seen that the performance does not change.

5.3. *Comparison of Proposed and CICIDS2017 Dataset with CS Method*

When calculating the performance, Table 5 shows the false positive rate and then compares the accuracy of these five machine learning classifiers. The accuracy of both datasets used the correlation-based feature selection method (CS) to calculate the performance of accuracy in the comparison. It proved that the false positive rate in both attacks of the Proposed Dataset is significantly lower when compared to CICIDS2017. The minimum accuracy of CICIDS2017 is 65.6% in Naïve Bayes (NB) and 64.2% in Bayes Net (BN), respectively, for PortScan attacks. In DoS attacks, the minimum accuracy of the proposed is 88.58% in Naive Bayes (NB). It founds that the accuracy of proposed Dataset is above 99% in both attacks.

Classifiers	Proposed				CICIDS2017			
	FPR (DoS)	FPR (PScan)	Acc (DoS)	Acc (PScan)	FPR (DoS)	FPR (PScan)	Acc (DoS)	Acc (PScan)
LG	0.1%	2.3%	99.9%	99.54%	27.3%	10.9%	82.48%	91.21%
NB	0.4%	0%	99.5%	99.54%	2.6%	42.9%	88.58%	65.6%
BN	0%	0%	99.99%	99.99%	0.9%	0.4%	98.26%	64.2%
J48	0.1%	0%	99.9%	99.99%	0.8%	0.2%	99.21%	99.77%
RT	0%	0%	99.99%	99.99%	0.8%	0.2%	99.22%	99.77%

Table 5. Comparison of FPR and Accuracy with CS Method

5.4. *Comparison of Proposed and CICIDS2017 Dataset with CA Method*

In CICIDS2017 dataset, the boundary value of the features obtained with the correlation Attribute (CA) method from the 77 full features without redundant feature show in Table 6, and 28 features are listed. From the 16 features in the proposed dataset, the eleven features acquired by setting and calculating the average boundary value of the solution obtained using the CA method. In determining the boundary value in the Proposed and CICIDS2017 Dataset, the average value calculates by adding the train of the features of the two datasets. The boundary value is set to 0.152955 for DoS and 0.07995 for PortScan attack.

Table 7 shows the performance of the false-positive rate and accuracy of the proposed and CICIDS2017 with percentages. If there were to express the FPR for each classifier for both datasets, it can see that J48 is 0.1% the same and the accuracy is almost the same. It can see that the NB classifier has an FPR of 9.8% in CICIDS2017, while the proposed one has only 0.4%. The RT classifier has an FPR of 0.2% in CICIDS2017 and 2.6% in the proposed. The LG and NB classifiers have FPR 1.1% and 1.5% in CICIDS2017, while the proposed has only 0.5% and 0.2%, respectively. In the proposed dataset, the accuracy of the RT classifier is close to 98%, and the other classifiers are above 99%. In the CICIDS2017 dataset, the accuracy is above 99% in J48 and RT; BN and LG have 98%, while the NB classifier has more than 79%, which shows significantly less accuracy.

No	Feature Code	Features	No	Feature Code	Features
1	13	Bwd Packet Length Mean	15	41	Average Packet Size
2	43	Avg Bwd Segment Size	16	18	Flow IAT Std
3	14	Bwd Packet Length Std	17	21	Fwd IAT Total
4	11	Bwd Packet Length Max	18	2	Flow Duration
5	23	Fwd IAT Std	19	12	Bwd Packet Length
6	38	Packet Length Std	20	22	Min
7	64	Idle Max	21	35	Fwd IAT Mean
8	62	Idle Mean	22	17	Min Packet Length
9	24	Fwd IAT Max	23	28	Flow IAT Mean
10	19	Flow IAT Max	24	1	Bwd IAT Std
11	65	Idle Min	25	29	Destination Port
12	36	Max Packet Length	26	40	Bwd IAT Max
13	37	Packet Length Mean	27	8	Down/Up Ration
14	39	Packet Length Variance	28	33	Fwd Packet Length Min Fwd Packets/s

Table 6. CA Method Result of CICIDS2017 Dataset

Detection Classifiers	CICIDS2017 CA Method			Proposed CA Method		
	Features - Code	FPR	Acc	Features No:	FPR	Acc
LG	1,2,8,11,12, 13,14,17,18,19, 21,22,23,24,28, 29,33,35,36,37, 38,39,40,41,43, 62,64,65	1.1%	98.83%	1,2,3, 4,7,8,10, 11,13,14,15	0.5%	99.5%
NB		9.8%	79.55%		0.4%	99.5%
BN		1.5%	98.18%		0.2%	99.8%
J48		0.1%	99.83%		0.1%	99.9%
RT		0.2%	99.98%		2.6%	97.91%

Table 7. Performance Comparison of DoS Attack with CA Method

5.5. *Discussion*

The proposed dataset creates sixteen features, and the goodness of these features proves the performance as a false-positive rate (FPR) and accuracy with six machine learning classifiers. A high false positive rate is not an attack, but an alert, so the organization may be affected by not being aware of the intruder's attack. Therefore, the proposed system sees the reduction of the false-positive.

In CICIDS2017, considering and not considering the thirteen flag features achieves the same performance in a DoS attack. Table 4 (a and b) shows that there is only a slight change in the PortScan attack and no impact on performance. If the attributes related to the thirteen flag features and the values are not considered, the calculation time and complexity time for performance is significantly reduced. The CICIDS2017 dataset using 28 features has an accuracy of 99.83%, and the accuracy of the proposed dataset using 11 features is 99.9%. It can be found in the J48 classifier of Table 7.

In [Kurniabudi (2019)], by setting the values of feature weight with CICIDS2017 dataset and using 4, 15, 22, 35, 52, 77 features and calculating performance with five classifiers, the highest accuracy is 99.87% in J48 classifier with 52 features, 99.86% in Random Forest classifier with 22 features, and 99.79% in Random Tree classifier with 15 features. When using many features, the execution time is significantly increased, but the accuracy is not seriously improved.

In addition, comparing the proposed dataset and the CICIDS2017 dataset, it wants to focus on the goodness of the features of the proposed dataset rather than CICIDS2017.

6. Conclusion

The main thing in the proposed system is to build a network testbed that includes a firewall, and IDS of the devices. A proposed dataset contains sixteen features from the network traffic of the testbed. The proposed system uses six machine learning classifiers, Logistic Regression, SVM, Naive Bayes, Bayes Net, J48 and Random Tree. It compares with CICIDS2017, excluding SVM, which has a long processing time, the system performance is proved using five classifiers. Comparing the proposed system and existing dataset CICIDS2017 is to compute the beneficial features and is applied a Correlation subset feature selection method (CS) and Correlation attribute method (CA) to know the impact of unnecessary features as calculation time and system resources. This work reduces the false positive rate and improve accuracy without affecting the network speed of the system performance.

Acknowledgments

The author is thankful to Prof: Dr. Khaing Khaing Wai, University of Computer Studies, Yangon (UCSY), for her crucial guidance, calmness, and motivation. The author also gratitude to her ICT Research Center (ICTRC) for assisting with the required resources.

Conflicts of interest

The authors reveal that they have no conflicts of interest.

References

[1] Abdallah, E.; Eleisah, W.; Otoom, A. F. (2022): Intrusion Detection Systems using Supervised Machine Learning Technoques: A Survey. The 13th International Conference on Ambient Systems, Networks and Technologies (ANT), pp. 205-212.
[2] Albarka, M.; et. al (2020): Network Intrusion Detection Using Wrapper-based Decision Tree for Feature Selection. IEEE INFOCOM, ACM, pp. 5-13.
[3] Alhomoud, A.; Munir, R.; Disso, J. P.; et al (2011): Performance Evaluation Study of Intrusion Detection Systems. Procedia Computer Science 5, published by Elsevier Ltd, pp. 173-180.
[4] Aljawarneh, S.; Yassein M. B.; Aljundi, M. (2017): An enhanced J48 classification algorithm for the anomaly intrusion detection systems. Cluster Comput., pp. 117.
[5] Applebaum, S.; Gaber, T.; Ahmed, A. (2021): Signature-based and Machine-Learning-based Web Application Firewalls: A Short Survey. Procedia Computer Science, pp. 359–367.
[6] Bijone, M. (2016): A Survey on Secure Network Intrusion Detection & Prevention Approaches. American Journal of Information System, Vol. 4, No.3, pp. 69-88.
[7] K. Bouzoubaa, K.; et.al (2021): Predicting DOS-DDOS Attacks: Review and Evaluation Study of Feature Selection Methods based on Wrapper Process. International Journal of Advanced Computer Science and Applications, **12**(5), pp. 131-145.
[8] Kamarudin, M. H.; Maple, C.; Wadson, T. (2019): Hybrid feature selection technique for intrusion detection system. Int. J. High Performance Computing and Networking, **13**(2), pp. 232-240.
[9] Kilincer, I. F.; Ertam, F.; et. al (2021): Machine learning methods for cyber security intrusion detection: Datasets and comparative study. Computer Network.
[10] Kshirsagar, D.; Kumar, S. (2021): An efficient feature reduction method for the detection of DoS attack. ICT Expert, 7, pp. 371-375.
[11] Kurniabudi; Stiawan, D.; et al. (2019): CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection. IEEE, 8.
[12] Liao, H.; Lin, C. R.; Lin, Y.; et al. (2013): Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications 36, pp. 16-24.
[13] Maseer, K. L.; et. al (2021): Benchmarking of Machine Learning for Anomaly Based Intrusion Detection Systems in the CICIDS2017 Dataset. Open Access Journal, pp. 22351-22370.
[14] Mauro, M. D.; Galatro, G.; et. al (2021): Supervised feature selection techniques in network intrusion detection: A critical review. Engineering Applications of Artificial Intelligence, 104216.
[15] Mohammadi, M. (2021): A comprehensive survey and taxonomy of the SVM-based intrusion detection systems. Journal of Network and Computer Application, 102983.
[16] Mukkamala, S.; Janoski, G.; Sung, A. (2016): Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithm. International Journal of Computer Applications, **150**(12).
[17] Panigrahi, R.; et. al (2021): Performance Assessment of Supervised Classifiers for Designing Intrusion Detection Systems: A Comprehensive Review and Recommendations for Future Research. Mathematics, 690.
[18] Pervez, P.S.; Farid, D. M. (2014)): Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs. The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014), Dec 2014.
[19] Tobi, A. M. A.; et. al (2020): Improving Intrusion Detection Model Prediction by Threshold Adaptation. Information, 10, 159.
[20] Thakkar, A.; Lohiya, R. (2020): A Review of the Advancement in Intrusion detection Datasets. Procedia Computer Science, 167, pp. 636-645.
[21] Valenza, F.; Cheminod, M. (2020): An Optimized Firewall Anomaly Resolution. Journal of Internet Services and Information Security (JISIS), pp. 22-37.
[22] Yang, L.; Shami, A. (2022): IDS-ML: An open source code for Intrusion Detection System development using Machine Learning. Software Impacts, 100446.
[23] Yi, H. H.; Aye, Z. M. (2019): Awareness of Policy Anomalies with Ruled-Based Firewall. ProMAC 2019, pp. 678-686.
[24] Yi, H. H.; Aye, Z. M. (2021): Performance Analysis of Traffic Classification with Machine Learning. International Conference on Information Technology and Electrical Engineering (ICITEE 2021), pp. 33-38.
[25] Zhang, L.; Huang, M. (2015): A Firewall Rules Optimized Model Based On Service-Grouping. 12th Web Information System and Application Conference, IEEE.

Authors Profile

Htay Htay Yi received the M. Sc degree for physics from Pathein University in 2001. The M. A. Sc degree of Master Applied Science accepted from University of Computer Studies, Yangon (UCSY) in 2003. She is an Associate Professor at the Department of Infromation and Communication Technology Research Center (ICTRC), Information and Communication Technology Training Institute (ICTTI). She certified CCNA and CCNA Security. She is an instructor of Cisco Networking Accademy of ICTTI. She researches interests include Networking, Network Security, Machine Learning, Cloud Computing, and Artifical Intelligence. She can be contacted at email: htayhtayyee@ictresearch.edu.mm.

Khaing Khaing Wai received the B.Sc. (Hons;), M.Sc. and M.Research. degrees in physics from Yangon University, Myanmar, in 1996, 1999 and 2000, respectively, and the Ph.D. degree in computer hardware technology from University of Computer Studies, Yangon, in 2005. She is currently Head of the Department of Information Technology Support and Maintenance at the University of Computer Studies, Yangon, Myanmar. She is also a Professor of cisco network lab in University of Computer Studies, Yangon. She can be contacted at email: khaingkhaingwai@ucsy.edu.mm.